**Target/Questions**

I want to use my data in order to

* estimate the fish population (i.e. understand that there is a deviation between the calculated fish number/average weight and the real ones)
* adjust feeding, based on the real fish population and the behavior of the fish
* predict a disease or another problem that will happen soon
* improve production efficiency by understanding the relationships between the parameters and their effect on the production result
* compare my performance to the one of other similar companies (benchmarking) and verify the validity of the findings of the previous bullet

The first 3 bullets are targets that have to do with the current situation and the management of the farm for the next period. The last two bullets is related both to current as well as long-term improvements

**Datasets**

There is a dataset for each running or closed cage. It has the attributes described in “Periodic Datasets” tab of “RequirementsGathering AquaSmart v0 1 - bm 20150220.xlsx”. There are three main groups of information

* KPIs that show the performance of the cage (output variables, blue cells) – **Group 0**
* Attributes of the fish population that exists in the cage (like hatchery, Fry CV, Stocking Year and Month, etc- gray cells). Those attributes do not change over time (with two exceptions, if the fish are vaccinated or not and the grading code) - **Group 1**
* Daily data of the fish. The dataset will contain either the information from the stocking of the fish until today or the information from the day the fish have been transferred to this cage until today). There will be a record for each day. There are 7 groups of information, shown in different colors (e.g. orange for the feeding-related data, light green for the net and density related, etc.) - **Group 2**

So the dataset is in the form master (KPIs and population attributes) – detail (daily records)

Each time it is submitted for analysis it will contain

* Data that are the same with the previous analysis (for example, the fish were stocked 1/1/2014 and the previous analysis was done on 15/1/2015, therefore the dataset contained the records from 1/1/2014 – 15/1/2015)
* New data that were produced from the previous analysis until today

Of course, for the new cages it will be completely new datasets

For each attribute

* There is a column that shows if it will be available for data mining at the global level.
* There are two groups of columns that show if the attribute exists currently in the data that the companies keep or it will exist in the future)
* There is a column for the companies to provide the importance they think this attribute has

**Analysis to be performed on the datasets**

We want to study

* The relationships between the parameters of **Group 2.** Some examples are provided in “interaction between parameters.xlsx”. For example, we want to know if the fish density influences the daily mortality
* The impact of the parameters of **Group 1** and **Group 2** to the KPIs of **Group 1**. We want to know the more important parameters, in order to focus our effort for improvement. I think there is a question here regarding the way we are going to consider the daily parameters. For example, we want to study the effect of the fish density to the daily mortality, the total mortality and the Economical FCR. How are we going to take into account the daily fish density regarding the LTD Econ. FCR? The input will be the daily values or just the average? In other words, we want to study the effect of a daily value to an LTD (Life to Date) Value
* The impact of the **change rate** of parameters of **Group 2** to other parameters of the same group or to the KPIs of **Group 0**. For example, does the mortality increase when temperature changes fast? So we are interested not only to the values but also to the rate they change.
* The effect of various parameters on the final distribution of the fish in harvest categories when harvested. Question: do we need to add another set of attributes to Group 1 that will be the distribution of the fish to harvest categories, for the closed units ??

Also, we want to identify missing fish or differences in the average weight (the first question) by looking at the data. For example, if there is a deviation between model and actual feeding for the last x days and if the behavior of the fish is that they are always at the bottom of the cage (not a real example) and if the fish density is OK and then most probably we have less fish than the calculated number. Input required: we need the users to provide a good set of rules on how to identify differences in the fish number or average weigh. What do they examine now, how they make conclusions.

And finally, we want the system to help users to identify problems in order to be able to react quickly. For example, if there is

* High deviation between model and actual feeding and/or
* high number of mortalities or quick increase of mortalities
* strange fish behavior
* big differences between estimated and actual average weight in samplings
* big number of holes the divers repair

Then it means that there is a problem. If for example the fish are losing appetite and the number of mortalities is increasing it is likely a disease is going to appear. Again, we need input here: we need to understand the way the users think. What they look first, what they do next, how they understand that something is happening.

**Questions to the users**

Regarding the datasets

1. How to register feed characteristics. (Fat% and Protein % but also we can use Fish meal and oil % that we know from the Feed companies)

* Fat % and Protein %
* Fish Meal % and Oil % per Kg
* Other

1. If we register feed characteristics, is it necessary to register feed energy class? (It’s difficult to know we will need feed analysis to expensive)
2. If the reply to question no 2 is yes, which are the energy classes?
3. For the mesh size, are we going to use the mesh size or total surface open (what is the meaning of total surface open?) (mesh size)
4. What are the exact parameters to be registered for net holes. Are number, size, depth necessary? (In our opinion we should register the total number of holes and also the sizes of each hole, and of course the depth (because is not the same have it in top of the net or in the middle of the net)
5. How the tidal is measured / registered? Is it a time value? Is it one value per day or multiple ones? (We never measure it, as Ardag will decide)
6. Does it make sense to register in sampling transactions the CV or standard deviation (we calculate in some sampling in others not, according the way of the sampling, but make sense introduce it)
7. What is the “Assessment” attribute? What are the possible values? We neither don’t use it.
8. Is there anything missing from the daily dataset? Shall the additional attributes provided by Mr. Anastasiadis be included in the datasets? No more. Yes, we should include, but he has to inform how to measure.

Regarding the decision making

1. What is the indication that something bad is happening in the production? For example,

* deviation between model and actual feeding
* high number of mortalities or quick inrease of mortalities
* fish behaviour
* big differences between estimated and actual average weight
* number of holes the divers repair

What are the check they do if the above is happening? We need some stories (pls see the example below) We make samplings biometrics to confirm the a.w. and confirm with the pathologist department that there aren’t any disease in the batch.

But also we need find the explanations, because there are a lot of cases that we can’t explain the losses at the end of the cage. And we need try to find a patron with:

* Storms
* Nº holes
* High mortality in the smallest fishes
* High mortality in short period of time
* Nº of net review by the divers
* Predators

1. What is the indication that the number of fish or the average weight in the cage are different from the calculated ones? What is the reasoning process now?

We need some stories that cover the main cases, for example

* There is a deviation between actual vs suggested in 10 days
* That means either I have losses or fish are going to be sick or the net is dirty
* First I check the divers comments. If no problems are reported I assume the net is clean and there are no holes
* Then I look at the oxygen. If it is OK, it is not the oxygen that causes the difference in the feeding
* Then I check the mortality rate. If it is increasing, I assume there is a disease
* If there is no disease, there are not the correct nº of fishes inside the cage or we have a deviation in the a.w.

In general, we need to understand the way the users examine their data. What they look first, what they do next, how they understand that something is happening.

**Questions to data mining experts**

* The datasets will be incremental. Each time they are submitted for analysis will contain the data that contained the previous time they were examined plus the new data. Is that a problem?
* We want to perform the analysis described in paragraph “**Analysis to be performed on the datasets**”. Is that possible?
* How the answers the users get will look like? I think this is the most important question. We need to provide some examples to make sure the analysis makes sense for them and helps them to answer the questions they want to answer.